world skills Russia

BRICS Business Council

机器学习和大数据
Machine Learning and Big Data

赛项技术规程
Technical Regulations

一、赛项介绍

竞赛名称：2019 喀山未来技能大赛专项赛机器学习和大数据赛项全国选拔赛

主办单位：金砖国家技能发展与技术创新大赛组委会

承办单位：金砖国家工商理事会（中方）技能发展工作组

赛项承办单位：重庆 Cloudera 大数据基地

赛项支持单位：北京嘉克新兴科技有限公司

北京企学研教育科技研究院

重庆翰海睿智大数据科技有限公司

二、竞赛内容

根据 2019 喀山未来技能大赛机器学习与大数据赛项操作技术方案核心要求，结合我方现有的技术平台，本选拔赛分为三个模块，竞赛总分为 100 分，每个模块占分比及比赛时长如表 1 所示。

表 1 竞赛模块

序号	比赛模块	时间	占分比
A	大数据技术	1.5h	40%
B	数据分析及可视化	1.5h	30%
C	机器学习	1.5h	30%
		总计：4.5h	总分：100 分

（一）大数据 （分数权重占 40%）

1. 竞赛技术平台

比赛采用基于 OpenStack 开发多节点大数据分布式实验平台

表 2 大数据技术平台

序号	功能/参数
1	采用 OpenStack 技术实现了硬件资源集中调度和管理
2	采用分布式技术搭建大数据竞赛环境

3	大数据实验集群的性能监控
4	实验环境快速部署、快速启动、快速恢复
5	实验环境相互独立互不干扰
6	支持 Linux 和 Windows 多种操作系统
7	支持大面积并发实验，稳定性强
8	支持横向拓展

2. 竞赛具体内容

赛项以大数据为核心内容和工作基础，重点考察参赛选手在 Hadoop 平台环境下，对于平台的搭建和基础组件的使用具体包括：基于大数据实验平台，完成 Hadoop 平台基本配置，数据抽取工具 Sqoop、分布式存储系统 HDFS、分布式计算框架 Yarn、流式数据仓库 Hive、实时分析查询引擎 Impala、Python 等开发语言工具和技术。

3.竞赛考核要点

表 3 大数据考核要点

模块	内容	配分
A	搭建大数据平台	10
A	使用 HDFS 文件系统	5
A	运行 Yarn 应用程序	5
A	用 Sqoop 导入数据	5
A	用 Hive 和 Impala 查询 HDFS	5
A	配置 HDFS 高可用性	5
A	创建 HDFS 快照	5

（二）模块 B：数据分析及可视化 （分数权重占 30%）

1. 竞赛技术平台

（1）比赛硬件采用独立电脑 PC 端。

（2）数据分析与可视化竞赛环境为 jupyter，采用 python 编程实现，技术平台如表 4 所示。

表 4 数据分析及可视化技术平台

序号	功能/参数
1	Python 中的 numpy 模块，支持大量的维度数组与矩阵运算，也针对数组运算提供大量的数学函数库
2	Python 中的 scipy 模块，支持处理插值、积分、优化、图像处理、常微分方程数值解的求解、信号处理等问题，也可用于有效计算 Numpy 矩阵，使 Numpy 和 Scipy 协同工作，高效解决问题
3	Python 中的 pandas 模块，提供了高效地操作大型数据集所需的工具，支持函数和方法快速便捷地处理数据
4	Python 中的 matplotlib 模块，是一个 Python 2D 绘图库，支持各种平台上以各种硬拷贝格式和交互式环境生成出具有出版品质的图形
5	Python 中的 seaborn 模块，是一个数据可视化库，提供更高级的 API 封装，支持更加方便灵活的应用
6	Python 中的 plotly 模块，支持制作交互式出版品质的图表，能够实现基本图表、统计图表、科学图表、财务图表、地图等多种类型图表
7	Python 中的 pyecharts 模块，用于生成 Echarts 图表。支持 30+ 种常见图表，多达 400+ 地图，为地理数据可视化提供强有力的支持

2. 竞赛具体内容

本赛项主要考核选手对数据分析流程的理解与各环节能力。结合现场提供的实验环境，对数据分析流程中的六个阶段依次实现，包括：数据整合、数据存

储、数据分析、数据可视化、报告撰写，最终形成一份完整的数据分析报告，作为最终作品提交。最后结合分析报告与竞赛过程情况，对选手进行综合评分。

3. 竞赛考核要点

表 5 数据分析及可视化考核要点

模块	内容	配分
B	数据整合： 多个不同格式数据源整合为完整单一数据源	6 分
B	数据探查与预处理； 查看源数据字段信息、样本分布情况、空值异常值处理等操作	6 分
B	数据分析： 主要内容为描述性统计分析，包括：最大值、最小值、均值、中位数、众数、方差、协方差等常用指标构建	6 分
B	数据可视化： 借助图形化手段，对数据探查和数据分析结果进行展示	6 分
B	报告撰写： 结合前面步骤分析结果，形成数据分析报告，提交竞赛作品	6 分

（三）模块 C：机器学习 （分数权重占 30%）

1. 竞赛技术平台

（1）比赛硬件采用独立电脑 PC 端。

（2）数据分析与可视化采用 python 编程实现，技术平台如表 4 所示。

表 6 机器学习技术平台

序号	功能/参数
1	Numpy：提供了存储单一数据类型的多维数组(ndarray)和矩阵（matrix）
2	scipy：其在 numpy 的基础上增加了众多的数学、科学以及工程计算中常用的模块，例如线性代数、常微分方程数值求解、信号处理、图像处理、稀疏矩阵等等
3	Pandas：提供高性能，易于使用的数据结构和数据分析工具
4	StatsModels：用于探索数据、估计统计模型、统计检验
5	Scikit-learn：提供经典的机器学习算法用于数据挖掘和数据分析
6	matplotlib：2D 绘图库，可绘制高质量的图片
7	Seaborn：是一个数据可视化库，提供更高级的 API 封装，支持更加方便灵活的应用
8	Plotly：支持制作交互式出版品质的图表，能够实现基本图表、统计图表、科学图表、财务图表、地图等多种类型图表
9	Pyecharts：用于生成 Echarts 图表。支持 30+ 种常见图表，多达 400+ 地图，为地理数据可视化提供强有力的支持

2. 竞赛具体内容

使用指定数据集，通过数据预览，探索式数据分析，缺失数据填补，删除关联特征以及派生新特征等方法，然后采用合适的机器学习算法构建模型对数据集进行训练和测试，并对模型的性能和预测能力进行测试。

分类模型评价指标：

- 准确率(Accuracy)

- 精确率(Precision)
- 召回率(Recall)
- F1值(F1 score)
- ROC和AUC

回归模型评价指标：

- 平均绝对误差
- 均方误差
- 决定系数

聚类模型评价指标：

- DB指数
- 轮廓系数
- Rand指数

3.竞赛考核要点

表 7 机器学习考核要点

模块	内容	配分
C	根据预测值和真实值的差值进行评分，结果越小越好	30 分